Network Science – Homework 1
Network Topology of Men's Single Tennis Matches

Umberto Michieli ID: 1150780
umberto.michieli@studenti.unipd.it

I. Introduction and Related Works

During the last decades Network Science field has been rediscovered and addressed as the "new science" [1], [2]. A lot of issues have been (re-)examined thanks to Network Science techniques, which are nowadays permeating the way we face the world as a unique interconnected component. The presence and the immediate availability of a huge amount of digital data describing every kind of network and the way in which its nodes interact, has made possible an interdisciplinary analysis of many large-scale systems.
Similar techniques have been recently applied also to professional sports, in order to discover complex interactions phenomena and universal rules which are almost invisible and difficult to recognize restricting the attention to small networks or to microscopic level. For example, complex-network analysis were conducted on soccer (e.g. in [3] and [4]), football ([5] and [6]), basket ([7] and [8]), baseball ([9]) and cricket ([10] and [11]), just to name a few.

In professional tennis as well, there are few studies examining how to map the matches into complex networks and then developing new ranking methods alternative to the ATP (Association of Tennis Professionals) official one (more on those methods will be likely discussed in the second homework).
The first work of this kind is represented by [12], where the authors explained the network generation and then they performed some simple analysis on single Grand Slams tournaments matches only (i.e. four tournaments each year: respectively *Australian Open, Roland Garros, Wimbledon* and *U.S. Open*). Then an important contribution was brought by [13], where a different network modeling is proposed and the PageRank algorithm is applied identifying *Jimmy Connors* as the most important single tennis player between 1968 and 2010.
More recent tennis-related complex network studies regard new ranking methods proposal and evaluations (see as reference [14], [15] and [16]), or are related to doubles matches [17] or to the gender and handedness effects in top ranking positions [18].

On the other side, however, in literature there is not an exhaustive and precise explanation about the network topology of the tennis matches graph. Moreover, some papers seems to be hasty in asserting a scale-free nature of the network with some inaccuracies. In this study we want to carefully analyze the resulting network structure when all the official single tennis matches are considered since the so-called "Open Era" to the end of August 2017 (i.e. from 1968 onward; the ATP organization, instead, was founded in 1972) and to state all its major properties which can be exploited for some interesting structural considerations, even not touched in the existing studies, and for further analysis. We also performed many computer simulations using Matlab.

II. Generation of Dataset and Network

We begin our discussion with the generation of the dataset: all men's tennis matches since 1968 are considered but the *Davis Cup* ones since they imply a completely different system for points' attribution and do not affect the total size of the network because they represent a minor contribution to the total number of matches. The data can be freely downloaded directly from the ATP website [19] and from other online repositories (like [20] for recent data and [21]) allowing to fix some inconsistencies in the official ones. Hence the first step to do is to merge the small datasets, provided on a yearly basis, in one only; this is a very delicate operation since we need to account for many format differences and bring them all back to a common language for the information's specification. Some preliminaries have been manually done using Microsoft Excel and subsequently through two Matlab scripts (`prepare_old_files.m` and the first parts of `read_tennis_files.m`). For the next considerations we decided to keep the following features of interest for each match: the tournament level, the tournament stage, the winner player and the loser player.
A brief excursion follows in order to explain those quantities. The tournament levels allows to identify the importance of a match, in fact ATP hosts tournaments of very different prizes (as regards both money and ATP ranking points assigned), which in increasing order of importance are: *ATP 250* tournaments, *ATP 500*, *Masters 1000*, the annual *ATP World Tour Finals* and the *Grand Slams* (different names were used in the past but similar considerations hold). The ATP points assigned to the winner of the tournament are respectively 250, 500, 1000, 1500 and 2000 and lower points are attributed to the players in proportion to the reached stage of the tournament (e.g round 128, round 64, up to semifinal and final); refer to Table II for a simplified overview of current points attribution distribution where points of qualified players are taken into account as last rounds of each entry.

With those considerations in mind it is possible to map the matches into many different network representations (in

	W	F	SF	QF	R16	R32	R64	R128
Grand Slam	2000	1200	720	360	180	90	45	10
ATP World Tour Finals	+500	+400	+200 points for each round robin match win					
Masters 1000	1000	600	360	180	90	45	25	15
ATP 500	500	300	180	90	45	20	10	
ATP 250	250	150	90	45	20	10	5	

Table I
ATP POINTS DISTRIBUTION. W=WINNER, F=FINALIST, SF=SEMI FINALIST, QF=QUARTER FINALIST, R=ROUND.

all of them, however, the nodes represent the players, hence they are homogeneous) and we are going to analyze three main scenarios which are summarized in Figure 1 matching the following descriptions:

1) **Direct graph representation:** in this model an edge exists from every loser to the winner, each link has a weight equal to the number of times the destination node won over the starting node. In case of multiple links the weights are just summed. Similar representations were adopted in [13] considering data up to 2010, in [12] considering data between 90s and 00s of male and female matches of Grand Slams only with different weights function, and in [14] with data of top-100 players only and different weights function. The obtained graph is not symmetric, not even if the respective unweighted version is considered. This representation is implemented in the Matlab script `read_tennis_files_wins.m`, which returns the adjacency matrix of this kind of network, and it is analyzed in the main script `tennis_wins.m`.
2) **Direct and symmetric graph representation:** this original proposal assumes the existence of a directed link from each couple of nodes which played at least one match against each other. The weights are the respective ATP points awarded by the two players; note that even the loosing player gets a non-negative points score. Also in this representation in case of multiple links the points are summed up. By construction, the network will be structurally symmetric but with possibly very different weights. This representation is implemented in the Matlab script `read_tennis_files.m` and it is analyzed in the main script `tennis.m`.
3) **Undirect (and unweighted) graph representation:** for some considerations it is useful to view the undirect and possibly unweighted version of the two representations just presented (which are naturally identical). Two players are connected through an undirected and unweighted edge if they play at least one match against each other, thus we obtain an undirected and symmetrical network. This representation is implemented on the fly in the script `tennis.m`.

The dataset is the largest possible since official tennis matches has been established and comprises of $N = 4245$ nodes (i.e. players) and a total of 151734 matches which leads to $L = 170168$ or 101436 links depending on the selected representation (larger number for the second representation). Notice that, as in many real networks, the matrix can still be

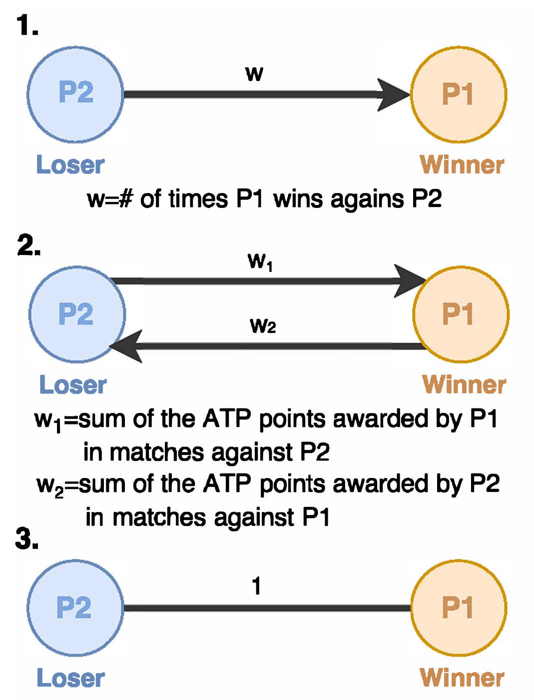

Figure 1. Three different network representations considered; each edge is associated to a respective weight briefly explained.

defined as sparse since it holds $L \ll L_{max} = \frac{N(N-1)}{2} = 9007890$ links, where L_{max} is the maximum number of links of a network with N nodes.
This large dataset will allow us to spot general trends and most competitive players overall; for more specific analysis it is enough just to restrict the attention to a smaller period of time (e.g. if we are interested in a specific player we should consider restricting our focus to his career epoch). Hence some results can be inherently biased toward the already retired players but in practice we will see that this does not always hold because of, for example, the increasing number of tournaments and of ATP points assigned each year.

III. RESULTS AND DISCUSSION

In this section we are going to explore the results of complex network techniques highlighting the properties and the underlying physical meaning. Moreover we will assert some comparisons among the different network representations to verify the common aspects through different views.

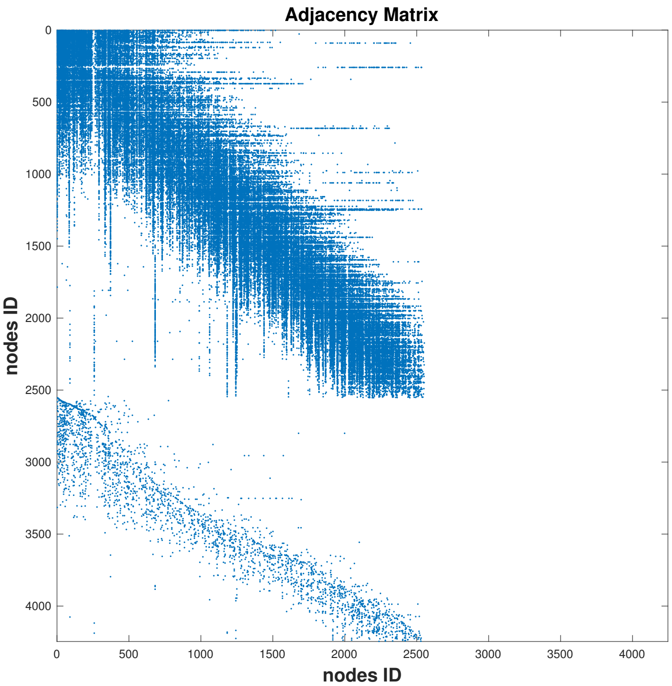

Figure 2. Adjacency matrix of the direct network representation.

A. Adjacency Matrix

In Figure 2 the adjacency matrix of the first representation of direct network is shown. The matrices in the other two scenarios can be simple found making the matrix symmetrical. The plot of the matrix has this shape because, by construction, in `read_tennis_file.m` first of all we have read the players which have won at least one match and after that we have considered the payers which figure only for lost matches. Thus, since there can't be any link between two always-loser players, the bottom-right part of the matrix is composed by all-zero entries. The bottom-left part of the matrix (and the respective up-right part if symmetric case is considered) do have a few points which are the matches lost by players who only have lost matches in the higher ATP tournaments (they surely have won some matches in minor ATP tournaments like *Challengers* or *Futures* in order for them to be admitted in the main draw of the most important ATP tournaments).
Moreover notice that the columns and the rows with a lot of non-zero entries are associated with players who have faced a lot of different players, thus usually they are players with a long-career and very *successful*, we should come back to this idea of evaluation of successful player in section III-E and in the second homework.

B. Network Visualization and Small World Property

In Figure 3 the visualization of the direct network is shown, since the other two representations could be easily derived starting from it.

By simply looking at the network topology we could already imagine that a giant component is present and that the small world property holds. By numerical evaluations,

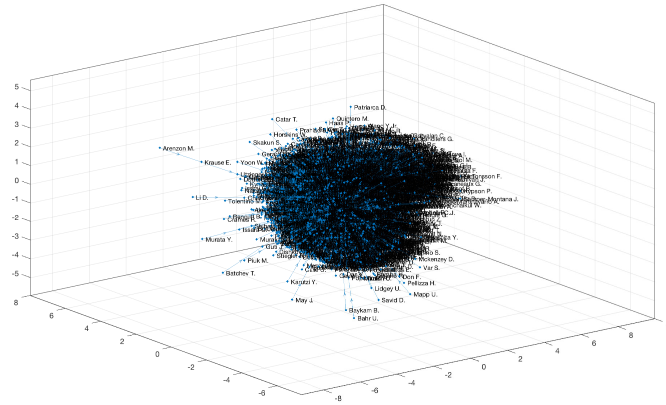

Figure 3. Direct network visualization.

indeed, we found that in the direct representation there is one giant component of size 2428 nodes and all the other components are unitary (in the undirect network there is just one component which contains all the nodes of the network). Defining the shortest path between any two node as the distance between those nodes we can derive the average distance of the direct graph defined as $\langle d \rangle = \frac{1}{N}\sum_{i=1}^{N} d_i$ with $d_i = \frac{1}{N}\sum_{j=1}^{N} \min(d_{ij})$ which lead to $\langle d \rangle \approx 3.48$ hops (and for the undirect network is $\langle d \rangle \approx 3.34$ hops). Moreover the diameter of the network, i.e. the maximum of all the shortest paths, is $diam = \max(\min(d_{ij})) = 10$ hops (and $diam = 8$ hops for the undirect network).
Hence, the network exhibits actually a *strong* small world property which leads to very short distances between any chosen pair of nodes. In order to better visualize it we could also look at the plots of the percentage of nodes within a considered directed hop distance as in Figure 4 and we realize that the worst cases are achieved only by a small fraction of nodes, thus reducing the variance of this metric. Notice that the blue curve for the first direct graph does not reach 100% because some of the nodes are disconnected from the giant component.

C. Degree Distributions

We are now interested in the evaluation of the degree distributions of the nodes in the network. We propose the distributions of in-degrees, of out-degrees and of total-degrees of the direct network in Figure 5. The values of $p_{in}(k)$, $p_{out}(k)$ and $p_{tot}(k)$ are the probabilities that a randomly picked node have k incoming, outgoing or total links; i.e. the fraction of nodes that have in-degree, out-degree or total-degree equal to k. The distributions follow a similar behavior and we can notice that are heavy-tailed, thus there are some hubs in the network, i.e. outliers at high values, and we are going to explore them in the next section.

The average in-degree and out-degree are $\langle k \rangle \approx 23.89$ edges, and the total average degree is then the double $\langle k \rangle \approx$

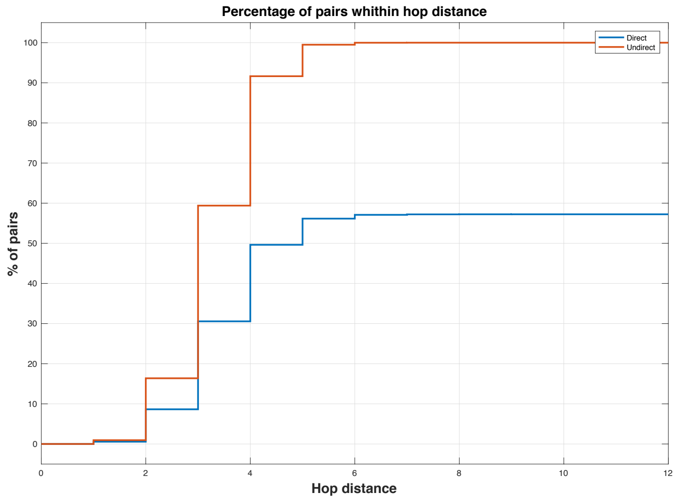

Figure 4. Percentage of pairs within hop distance for direct networks (i.e. representations 1 and 2 and actually for the second one the undirect case is equivalent), both are considered unweighted.

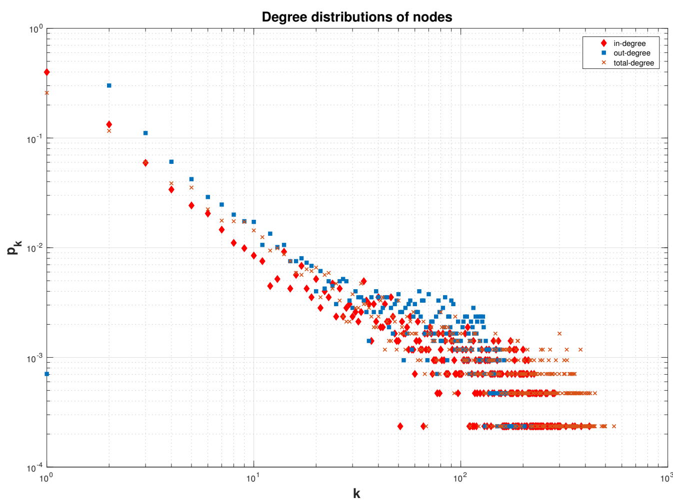

Figure 5. Direct network degree distributions for in, out and total links in log-log scale.

47.78 edges. As already stated, the network is fully connected, which makes sense being $\langle k \rangle$ such a high value.

The second moment for in-degree is $\langle k_{in}^2 \rangle \approx 3.31 \cdot 10^3$, for out-degree is $\langle k_{out}^2 \rangle \approx 2.01 \cdot 10^3$ and for total-degree is $\langle k_{tot}^2 \rangle \approx 1.01 \cdot 10^4$.

D. Hubs

We have already pointed out that the considered graphs present some hubs which are for interest in starting to determine the importance of a player in limitations to some specific metrics.

Indeed, the strongest players, identified as hubs, tend to play against a wide range of players: the *weak* ones, generally at the first stages of the tournaments (the top-players, as tournament's seeds, are facilitated in the first rounds when they are called to face qualified players, which are generally

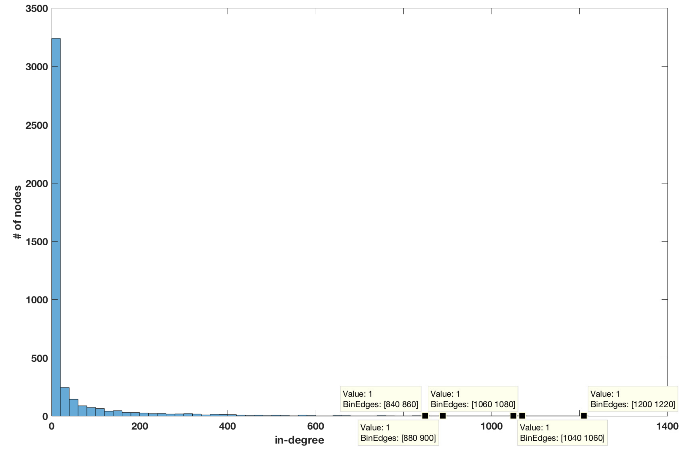

Figure 6. Histogram of occurrences of in-degree nodes, in the x-label the index of the players.

weak), and the *strong* ones at the last stages of the tournament, which are rarely reached by the weak players.

From the three typologies of network representation we can extract some useful information by simply looking at those hubs.

First of all, in Figure 6 we can see the histogram of occurrences of in-degree nodes in linear scale in order to be aware that the five highest-degree nodes are well spread apart from the majority of the other nodes which have much lower degrees.

Thus if we look at the first five in-degree biggest hubs we can determine the players who won more matches and the respective number of winnings, those are: *Jimmy Connors* (1219), *Roger Federer* (1076), *Ivan Lendl* (1047), *Guillermo Vilas* (892) and *John McEnroe* (840). If we compare with the ATP results archive we see slight variations, lower than 4%, due to *Davis Cup* matches.

Analogous procedures apply to the players who have lost the major number of matches: *Fabrice Santoro* (436), *Feliciano Lopez* (387), *Mikhail Youzhny* (379), *Guillermo Vilas* (373) and *John McEnroe* (351). Notice that here the numbers are quite different from before, because only the players who also win a lot of matches are guaranteed to play in the major tournaments, otherwise after a while you loose the right to play in.

Surely the results seen up to now tend to promote already retired players and/or with a long career. On the other side, if we think of applying similar techniques for the network with ATP ranking weights, we see quite different outcomes promoting current-time players because nowadays there are great players, of course, but also more tournaments which give even more ATP points than in the previous generations of tennis players. Thus the five players who gained the most ATP points are: *Roger Federer*, *Rafa Nadal*, *Novak Djokovic*, *Jimmy Connors* and *Ivan Lendl*.

Looking at the degree distributions in Figure 5 and at the results we are suggested to consider an underlying assumption:

the more connected athletes are and the most likely is to be best players. Most of the players have a small number of matches and then quit playing the major tournaments, on the contrary, there is a small group of top-players who perform many matches against weaker players and among themselves. This phenomenon is an observation of the *rich get richer* effect driven by the attractiveness of the high connected nodes as opponent for new-comers; an interpretation of the *richness* that the players achieve could be their gain of some sort of "experience" during the matches of their carreer, as already pointed out in [12].

E. Considerations on Network Nature

Then we came to the most critical point in the analysis already present in literature on this topic, i.e. the scale-free nature of the network. A first-step analysis about the network nature can be done by plotting again the degree distribution (e.g. the in-degree) and trying to fit it with some typical network distributions. The results are shown in Figure 7 where we can appreciate the differences among them. The respective parametric formulations of distributions together with fitting parameters and the coefficient of determination R^2 are reported in Table II. As already noticeable by the presence of hubs, the network cannot clearly follow a random model (the poisson one) and some heavy-tailed distributions need to be considered. The power-law and the Lévy distributions are the two models performing better on the raw data, thus the considered network exhibits many properties typical of scale-free networks.
Assuming the network as power-law, we can measure the scale-free parameter γ: it is found to be $\gamma_{in} \approx 1.66$ for in-degrees, $\gamma_{out} \approx 2.12$ for out-degrees and $\gamma_{\text{undirect}} \approx 1.27$ for the undirect network, values which are consistent with the ones found in literature in [13] and [12].

The aforementioned results need to be taken with some cautions; the network characteristics are quite similar to scale-free networks (they are even more similar when we restrict our interest to top-players and/or to top-tournaments only) and scaling behavior is also suggested by the structural *preferential attachment* of new players who generally tend to connect to an existing player with a probability proportional to the degree of such node. However if we try to reduce the noise around the outliers, e.g. by considering the cumulative degree distribution or a log-binning of the data, we can see that the network is not a pure scale-free model and it would make sense to limit the intervals introducing cutoff values k_{min} and k_{max}, as carefully proved in [22] in a general setting. Moreover, as suggested in [23] we should not rely on the R^2 parameter, since it is proved to achieve very high values also for non scale-free networks. Summing up: the fit on raw data based on R^2 is source of many errors in current literature about network analysis.

Hence, the Complementary Cumulative Distribution Function (CCDF) should be considered and it is plotted in Figure 8; here the issue of the plateau corresponding to values occurring once has been solved and we can confirm the previous considerations since power-law networks would describe a straight line in a log-log plot.

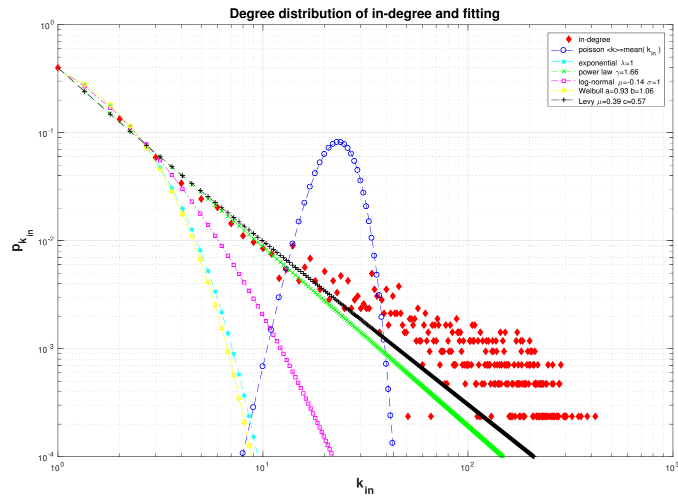

Figure 7. Fitting trials of raw data of in-degrees, log-log scale.

More specifically, replicating the fitting on the cumulative distribution, the network can be categorized as a power-law with an exponential cut-off, i.e. $p_k \propto x^{-\alpha} e^{-\frac{k}{\beta}}$: indeed we can observe in Figure 8 that the curve starts out as a power law and ends up as an exponential.
This result confirms that R^2 on raw data cannot be trusted and that, maybe, often is just worth noticing that the distribution has a heavy tail instead of asserting immediately its scale-free nature, which is a very widespread practice (many interesting considerations about networks and fitting techniques can be read in [23]).
The fact that the complete network is not scale-free is a quite surprising result, although very clear from the CCDF plot, because all other related studies on tennis network are affirming the scale-free nature and are deriving from there the interesting properties of the network, which is not precisely correct.
Some power-law networks with coefficient lower than 2 or not scale-free networks at all have gained a lot of attention in recent literature, for example in [24] and [25], because those models have been discovered in many real scenarios where the number of new links generally grows faster than the number of new nodes, which is precisely our situation. The number of annual tennis matches is nowadays very big and therefore the probability of a new connection is much more frequent than the new players who join the most important ATP tour. Thus it may seem that the network should evolve toward a non-sparse adjacency matrix, which is still not the case because of structural constraints: is very unlikely that some pairs of players will face against each other (because of players' retirement from the tour or players which are strongly far apart in ranking) but on the other hand the hubs role of long-career players will be even reinforced.

	pmf or PDF	Numerical parameters		R^2
Poisson	$p_k = e^{-\lambda}\frac{\lambda^k}{k!}$	$\lambda = \langle k \rangle$		
Exponential	$p_k = \lambda e^{-\lambda k}$	$\lambda = 1$		0.9879
Power-law	$p_k \propto k^{-\gamma}$	$\gamma = 1.66$		0.998
Log-normal	$p_k = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln(x)-\mu)^2}{2\sigma^2}}$	$\mu = -0.14$	$\sigma = 1$	0.9938
Weibull	$p_k = abx^{(b-1)}e^{-a*x^b}$	$a = 0.93$	$b = 1.06$	0.9866
Lévy	$p_k = \sqrt{\frac{c}{2\pi}}\frac{e^{-\frac{c}{2(x-\mu)}}}{(x-\mu)^{3/2}}$	$\mu = 0.39$	$c = 0.57$	0.9979

Table II
FITTING DISTRIBUTION APPLIED TO IN-DEGREE RAW DATA.

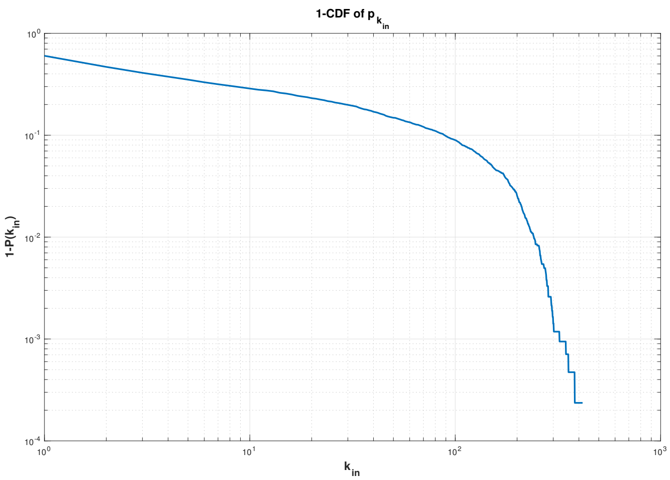

Figure 8. Complementary Cumulative Distribution Function (CCDF) in log-log plot.

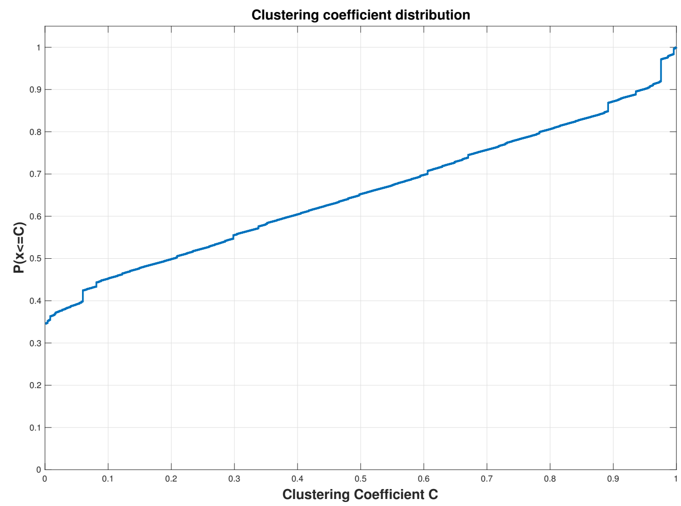

Figure 9. Clustering coefficient distribution of the undirect representation.

F. Clustering Coefficient

Another interesting property of small world network is the clustering coefficient C. For a node j its clustering coefficient, C_j, is a number belonging to $[0, 1]$ denoting how many links there are between its neighboring nodes normalized to the maximum possible number of links among them; more formally could be defined as $C_j = E_j/E_{j,\max}$ where E_j is the number of edges between nodes in the neighborhood of j ($\mathcal{N}_j$) and $E_{j,\max}$ is its maximum value.
For example, in undirect networks it holds $0 \leq E_j \leq E_{j,\max} = |\mathcal{N}_j|(|\mathcal{N}_j| - 1)/2$. Finally, the general clustering coefficient is expressed as the average over all the N players: $C = \frac{1}{N}\sum_{j=1}^{N} C_j$.
In the undirect representation of the network we have found $C = 0.07$, which is coherent as order of magnitude with the values found in [12]; as a term of comparison, in Figure 9 is reported the clustering coefficient distribution, i.e. the fraction of nodes having clustering coefficient lower than c.

G. Degree Correlation

Another interesting metric on the network point of view is the correlation between the degrees of nodes. Degree correlation can be expressed in many ways, in the following we are going to examine two of them: the degree correlation matrix and the assortativity coefficient.
The degree correlation matrix, E, is a square matrix of size the maximum value of the degrees, $\max(k)$. Each entry e_{ij} is the probability of finding a link between two nodes of degree i and j. The matrix is represented as heatmap for the undirect and unweighted network in Figures 10 and 11 (the color differences can be spotted by zooming at them) but we certainly notice how difficult is the visual inspection and it is actually impossible to infer anything, in addition to that it is also high computationally demanding.

For the reasons yet mentioned, usually the Pearson assortativity coefficient is used as principal investigation method and is defined as:

$$r = \frac{\sum_{h=\min(k)}^{\max(k)} \sum_{f=\min(k)}^{\max(k)} hf(e_{hf} - q_h q_f)}{\sigma^2}$$

where q_h is the probability of finding a degree-h node at the end of a randomly picked link, σ^2 is the variance of the degrees and can be proved to be the maximum of the numerator, thus $r \in [-1, 1]$. Computing this parameter for the undirect and unweighted network we found $r = -0.0076$ which means a slightly assortative network and actually almost

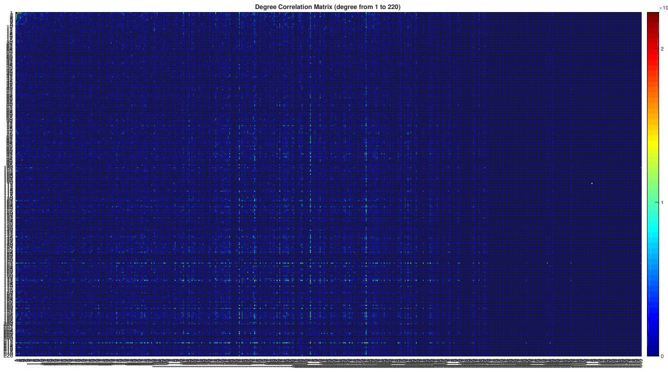

Figure 10. Correlation matrix of degrees from 1 to 220 in the ordinates.

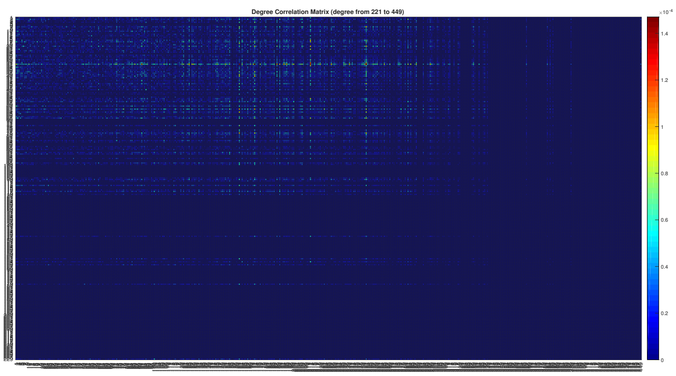

Figure 11. Correlation matrix of degrees from 221 to $\max(k) = 449$ in the ordinates.

neutral, probably because hubs and small degree nodes are likely to play against in the first stages of the tournaments but also hubs tend to play against themselves in the final rounds, thus no strong pattern exists.

H. Robustness to Failures

For a more complete characterization of the network structure we could be interested in its robustness to failures, i.e. the nodes removal from the network. In our network a player could be disqualified for doping or other reasons, or we could need to consider just a subset of players or matches (restricting by nationality, left or right handedness, height threshold, tournament level and so on). We want to determine the robustness of the network in terms of nodes connected to the giant component when $f\%$ of its nodes has been removed. In the following we are just considering *random* removals and *attack-based* removals since all the other are mainly application-driven and can be done with a small effort manipulating the dataset as desired. In the first scenario considered the nodes are removed entirely at random while in the second scenario the highest hub in the network is removed at each step. We introduce the probability that a random node belongs to the giant component after that $f\%$ of nodes have been removed as $P_\infty(f)$ and we can look at the relative size of the giant component: $P_\infty(f)/P_\infty(0)$, where $P_\infty(0)$ represents the best case of no removals thus the ratio belongs to $[0, 1]$. The plot of such ratio for the undirect graph is shown in Figure 12 and we recognize in our network the high robustness typical to well-connected and scale-free graphs. The black line corresponds to $1 - f$ and it is an upper limit since for sure we have removed $f\%$ of nodes from the network (and then also from the giant component).

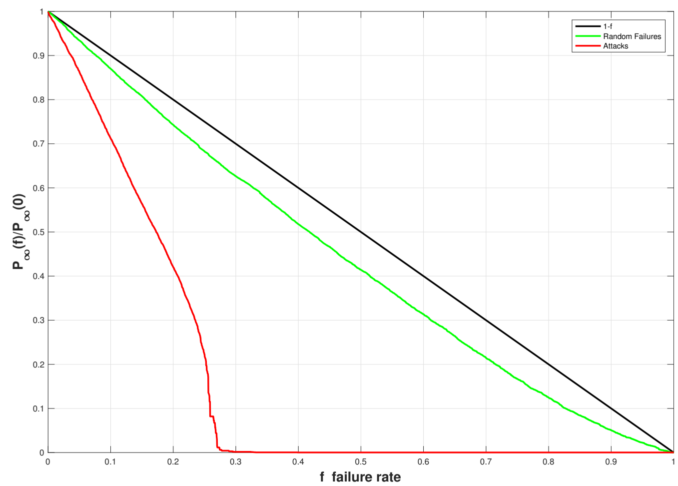

Figure 12. Robustness of the undirect network after $f\%$ of nodes removals.

IV. Conclusions and Future Work

In this study we have shown how it is possible to map the ATP single tennis matches into different graph representations; then we evaluated some metrics typical of those networks and we compared the results with the existing literature compensating for the lack of structural analysis of the network. Our future interest is addressed to exploit those network representations for more practical considerations: for example we would like to examine some suitable ranking methodologies different from the official ATP system and eventually to compare them, as well as to discuss the possibility of some sort of new-links prediction (i.e. matches between players who have never faced) and winner prediction power too.

References

[1] A. L. Barabási, "Linked: The new science of networks," 2003.

[2] T. G. Lewis, *Network science: Theory and applications*. John Wiley & Sons, 2011.

[3] T. U. Grund, "Network structure and team performance: The case of English Premier League soccer teams," *Social Networks*, vol. 34, no. 4, pp. 682–690, 2012.

[4] R. N. Onody and P. A. de Castro, "Complex network study of Brazilian soccer players," *Physical Review E*, vol. 70, no. 3, p. 037103, 2004.

[5] R. Trequattrini, R. Lombardi, and M. Battista, "Network analysis and football team performance: a first application," *Team Performance Management*, vol. 21, no. 1/2, pp. 85–110, 2015.

[6] J. Gama, P. Passos, K. Davids, H. Relvas, J. Ribeiro, V. Vaz, and G. Dias, "Network analysis and intra-team activity in attacking phases of professional football," *International Journal of Performance Analysis in Sport*, vol. 14, no. 3, pp. 692–708, 2014.

[7] F. M. Clemente, F. M. L. Martins, D. Kalamaras, and R. S. Mendes, "Network analysis in basketball: Inspecting the prominent players using centrality metrics," *Journal of Physical Education and Sport*, vol. 15, no. 2, p. 212, 2015.

[8] P. O. Vaz de Melo, V. A. Almeida, and A. A. Loureiro, "Can complex network metrics predict the behavior of NBA teams?," in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 695–703, ACM, 2008.

[9] S. Saavedra, S. Powers, T. McCotter, M. A. Porter, and P. J. Mucha, "Mutually-antagonistic interactions in baseball networks," *Physica A: Statistical Mechanics and its Applications*, vol. 389, no. 5, pp. 1131–1141, 2010.

[10] S. Mukherjee, "Identifying the greatest team and captain—A complex network approach to cricket matches," *Physica A: Statistical Mechanics and its Applications*, vol. 391, no. 23, pp. 6066–6076, 2012.

[11] S. Mukherjee, "Quantifying individual performance in Cricket—A network analysis of batsmen and bowlers," *Physica A: Statistical Mechanics and its Applications*, vol. 393, pp. 624–637, 2014.

[12] H. Situngkir, "Small world network of athletes: Graph representation of the world professional tennis player," 2007.

[13] F. Radicchi, "Who is the best player ever? A complex network analysis of the history of professional tennis," *PloS one*, vol. 6, no. 2, p. e17249, 2011.

[14] D. Aparício, P. Ribeiro, and F. Silva, "A subgraph-based ranking system for professional tennis players," in *Complex Networks VII*, pp. 159–171, Springer, 2016.

[15] A. D. Spanias and B. W. Knottenbelt, "Tennis player ranking using quantitative models," 2013.

[16] J. Maquirriain, "Analysis of tennis champions' career: how did top-ranked players perform the previous years?," *SpringerPlus*, vol. 3, no. 1, p. 504, 2014.

[17] K. Breznik, "Revealing the best doubles teams and players in tennis history," *International Journal of Performance Analysis in Sport*, vol. 15, no. 3, pp. 1213–1226, 2015.

[18] K. Breznik, "On the gender effects of handedness in professional tennis," *Journal of sports science & medicine*, vol. 12, no. 2, p. 346, 2013.

[19] "Results archive of ATP World Tour offciial matches since 1968: `http://www.atpworldtour.com/en/scores/results-archive`,"

[20] "Historical Tennis Results and Betting Odds Data since 2000: `http://www.tennis-data.co.uk/data.php`,"

[21] J. Sackmann, "Tennis Abstract, webpage of tennis analytics: `http://www.tennisabstract.com/blog/`,"

[22] C. I. Del Genio, T. Gross, and K. E. Bassler, "All scale-free networks are sparse," *Physical review letters*, vol. 107, no. 17, p. 178701, 2011.

[23] A. Clauset, C. R. Shalizi, and M. E. Newman, "Power-law distributions in empirical data," *SIAM review*, vol. 51, no. 4, pp. 661–703, 2009.

[24] J. Ugander, B. Karrer, L. Backstrom, and C. Marlow, "The anatomy of the facebook social graph," *arXiv preprint arXiv:1111.4503*, 2011.

[25] H. Seyed-Allaei, G. Bianconi, and M. Marsili, "Scale-free networks with an exponent less than two," *Physical Review E*, vol. 73, no. 4, p. 046113, 2006.